ระบบจัดกลุ่มข้อความอัตโนมัติโดยใช้เทคนิคการเรียนของเครื่อง
Automatic Text Clustering System
using Machine learning Techniques

พีระพัชร์ โกมลรุจินันท์ (Peeraphat Komolruchinonth)[1] วิศรุต ยาวุธ (Wisarut Yawut)[2]
และ ผศ.ดร วันทนี ประจวบศุภกิจ(Asst.Prof.Dr.Wanthanee Prachuabsupakij)[3]
ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีและการจัดการอุตสาหกรรม
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
[1]5706021630071@fitm.kmutnb.ac.th, [2]5706021610118@fitm.kmutnb.ac.th, , [3] wanthanee.p@fitm.kmutnb.ac.th

บทคัดย่อ

บทความวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาระบบการจัดกลุ่มข้อความอัตโนมัติโดยใช้เทคนิคการเรียนรู้ของเครื่อง กรณีศึกษาเป็นข้อความเกี่ยวกับสถานที่ท่องเที่ยวในประเทศไทยบนทวิตเตอร์ หรือเรียกว่า ทวิต ระบบที่พัฒนาขึ้นนี้ถูกพัฒนาในลักษณะเว็บแอพพลิเคชั่น ที่ดึงทวิตภาษาอังฤษจากทวิตเตอร์โดยใช้ Twitter API ซึ่งมีจำนวนทั้งสิ้น 75,319 ทวิต (ตัวอย่าง) หลังจากนั้นนำข้อความที่ได้มาผ่านกระบวนการทำเหมืองข้อความเพื่อให้ได้เวคเตอร์ (Vector) โดยใช้เทคนิคการเรียนรู้ของเครื่อง และจัดกลุ่มด้วยขั้นตอนวิธี K-means และพิจารณาจำนวนกลุ่มที่เหมาะสมจากทฤษฎีเอลโบล์ (Elbow) ซึ่งให้จำนวนกลุ่มทั้งสิ้น 7 กลุ่ม จำนวนข้อความในแต่ละกลุ่มมีดังนี้ กลุ่มที่ 1 ได้ 46,993 ตัวอย่าง กลุ่มที่ 2 ได้ 2,146 ตัวอย่าง กลุ่มที่ 3 ได้ 9,516 ตัวอย่าง กลุ่มที่ 4 ได้ 1,593 ตัวอย่าง กลุ่มที่ 5 ได้ 1,595 ตัวอย่าง กลุ่มที่ 6 ได้ 4,261 ตัวอย่าง กลุ่มที่ 7 ได้ 11,215 ตัวอย่าง หลังจากนั้นผลการจัดกลุ่มข้อความจะถูกนำมาแสดงผลผ่านทางเว็บแอพพลิเคชั่น เพื่อสามารถนำไปใช้ประโยชน์ในด้านต่าง ๆ ต่อไป

คำสำคัญ: จัดกลุ่มข้อความ, การเรียนรู้ของเครื่อง, เคมีน,ทวิตเตอร์, การทำเหมืองสื่อทางสังคม

Abstract

The aim of this paper is to develop Automatic Text Clustering System using Machine learning techniques. This text is streamed from Twitter in terms of Tourist Attraction of Thailand using Twitter API including 75,319 instances. The system is web-based application. Then, the instance is preprocessed using text mining to generate the feature vectors for each instance. After that, K-means is used as the clustering algorithm and we use elbow to help finding the appropriate number of clusters, which give 7 clusters including cluster 1 have 46,993 instances, cluster 2 have 2,146 instances, cluster 3 have 9,516 instances, cluster 4 have 1,593 instances, cluster 5 have 1,595 instances, cluster 6 have 4,261 instances and cluster 7 have 11,215 instances. Moreover, all results are reported on a web application that can be used for various domains.

Keyword: Text Clustering, Machine Learning, K-means, Twitter, Social Media Mining.